Checking Normality

Author: William Morgan, The College of Wooster

Focus: How to use the “three-prong” approach to check for normality

Overview: This lesson introduces students to the three-pronged approach for checking normality. During the lesson, students visualize the distribution of a continuous numerical variable in a biological data set using histograms and normal quantile plots, and perform a goodness-of-fit test. By comparing their outputs to a known normally distributed variable, students are able to correctly interpret the results of this three-pronged approach when working with data that deviates from a normal distribution.

Learning objectives:

By the end of this lesson, a student will be able to…

- Construct a histogram and identify data that deviates from a normal distribution.
- Script the necessary commands to construct a normal quantile plot.
- Correctly interpret a normal quantile plot identify data that deviates from a normal distribution.
- Script the necessary commands to perform the Shapiro-Wilk test and correctly interpret the results to check for normality.
- Identify variables that do not meet the assumptions of the t-test.

Lesson sequence: Provide a numbered, ordered list of the activities within your swirl lesson. This list can be taken from step 4 in your initial lesson design, with any modifications that were introduced.

1. Introduction to the biological data set and the research question for this lesson √
2. Assignment Problem 13-21 from The Analysis of Biological Data, 2nd ed. by Whitlock & Schluter
3. This requires a paired t-test
4. Students must select the appropriate statistical test √
5. Students identify the assumptions of the t-test √
6. Students review the concept of a normal distribution and standard deviation
7. Students construct a histogram using ggplot
8. Students compare the histogram to the normal distribution and indicate if it deviates from the expectations of a normal distribution

9. Students are introduced to the normal quantile plot starting with the two axes.
10. The normal distribution is mapped onto the Y-axis
11. Half of the normal distribution is below the mean/median
12. 95% of the distribution is within ~2 SD of the mean
13. 2.5% of the distribution is less than mean - ~2*SD & 2.5% is more than mean + ~2*SD
14. The data distribution is mapped onto the X-axis (see here for specifics)
15. Each data point is plotted on the X-axis
16. The data point is then raised up on the Y-axis based on its quantile: e.g., if 20 data points, the lowest value is at the 2.5% quantile, which we expect to be -2 SD from the mean; the value ranked 10th of 20 is at the ~50% quantile, which we expect to be at the mean (0 SD from the mean); the value ranked 20 of 20 is at the 97.5% quantile and is expected 2 SD from the mean.
17. The normal quantile plot of a variable exhibiting a normal distribution is examined and described.
18. Students randomly sample a normal distribution using rnorm()
19. Students make a normal quantile plot using the qqplot() geom with ggplot()
20. Students construct a normal quantile plot of the data, compare it to the previous plot, and indicate how it differs from the plot of an ideal normal distribution.
21. Students are introduced to the command to run a Shapiro-Wilk test.
22. Students perform the test on an idealized data set with a normal distribution and then with their biological data.
23. Students use the three-pronged approach to complete Assignment Problem 13.21 in The Analysis of Biological Data